



For reprint orders, please contact:
reprints@futuremedicine.com

Ultra-high-throughput sequencing, microarray-based genomic selection and pharmacogenomics



Gary Hardiman

BIOGEM

and,

University of California San Diego, Department of Medicine, La Jolla, CA 92093-0724, USA
Tel.: +1 858 822 3792;
Fax: +1 858 822 6430;
E-mail: ghardiman@ucsd.edu

'Appropriate retraining of medical personnel in genomic medicine will be an *a priori* requirement so that they are better equipped to counsel and treat patients presenting with the awareness of possessing genetic aberrations.'

The adoption of a novel technology by the scientific community typically heralds the start of a new era, where assay throughput is increased by an order of magnitude, and accompanied by an exponential reduction in cost compared with pre-existing approaches. The economic benefit and efficacy of emerging technologies are realized by process-miniaturization combined with the multiplexing of millions of reactions. A classic example is the DNA microarray or biochip, which became widely available over a decade ago and was subsequently applied to studies encompassing many different organisms, cell types and scientific disciplines [1,2]. Biochips evolved steadily over time from archetypal in-house cDNA arrays to robust commercial oligonucleotide platforms, a progression that was marked by migration to higher-density biochips with increasing content [3,4]. Often, appreciation of the benefits and long-term impact of emerging technologies is not immediate, since initial use is frequently restricted to developers. The impact of the technology and its long-term value is not felt until its wider acceptance by scientists who apply the technology to ask key fundamental research questions and, in the process, enhance its scope and utility [5].

Currently ultra-high-throughput DNA sequencing is transitioning from development to widespread use with several academic and commercial efforts aimed at developing ultra-low-cost sequencing [6-7]. The ultimate aim of these technologies is to reduce the costs of DNA sequencing by several orders of magnitude, and in essence facilitate the establishment

of a genome sequencing center in every laboratory [6]. Incorporation of next-generation sequencing technologies into biomedical research and drug discovery programs is currently a key issue at many academic institutes and pharmaceutical companies. As the technology is nascent, many changes can be expected in the next few years, mirroring the early development of microarray technology, where early adopters were forced to build on their initial investment and commit additional resources in the form of equipment upgrades or hardware replacement to remain current with the technology. This adds a layer of complexity to the choice of the most appropriate platform. Factors including cost, ease of use, versatility, peer review in the form of published data, platform stability and long-term technical support will guide the selection process. Data management issues, specifically the annotation, storage and retrieval, will pose formidable challenges. As was the case with microarrays, improved analytical tools will likely emerge over time, permitting additional analyses and enhanced information mining from raw data sets.

'Often, appreciation of the benefits and long-term impact of emerging technologies is not immediate, since initial use is frequently restricted to developers.'

A new utility for DNA microarrays has been described in several recent publications [8-11]. High-density oligonucleotide microarrays can be repurposed as hybrid-selection matrices to capture defined genomic fragments as substrates for sequencing. This represents a paradigm shift from conventional array-based approaches where DNA hybridization to a cognate probe generates a coordinate signal and the intensity is translated into biological information [9].

Traditional DNA sequencing

In 1975, Sanger and Coulson published the seminal 'plus-minus method' of DNA sequencing, which employs DNA polymerase to catalyze the reaction [12], and sequenced the

5375-nucleotide (nt) genome of bacteriophage phi X 174 [13]. Then, 2 years later, Maxam and Gilbert reported an alternate DNA-sequencing method based on the chemical modification of DNA and subsequent cleavage at specific bases [14]. Although both represented ground-breaking advances, the ‘chemical modification’ and ‘plus–minus’ methods lacked the efficiency of the ‘chain-termination’ method that was subsequently developed by Sanger and coworkers [15]. This approach utilized dideoxynucleotide triphosphates (ddNTPs) as DNA-chain terminators and afforded reductions in the amounts of radiolabeled DNA and other toxic reagents.

For the past three decades, the DNA-sequencing field has, consequently, been dominated by ‘Sanger’s chain-termination’ method.

‘One of the primary objectives of next-generation sequencing technologies has been to circumvent the cumbersome DNA cloning and library construction steps.’

Technological advances and major innovations in instrumentation have advanced the throughput of genome sequencing, reducing tedious and time-consuming missions to routine projects that are now accomplished in a matter of months. However, the costs associated with sequencing a single human genome using traditional Sanger sequencing remain elevated and are estimated in the region of US\$ 10–25 million [7].

Next-generation DNA sequencing

Traditional sequencing approaches require the fragmentation of large DNA polynucleotides into smaller pieces, followed by amplification and sequencing of the individual fragments, data quality control and, finally, computer-based assembly of contiguous sequences. When combined, these steps can take over 3 weeks to accomplish [6,7]. One of the primary objectives of next-generation sequencing technologies has been to circumvent the cumbersome DNA cloning and library construction steps.

Margulies *et al.*, described an early 454 Life Sciences (Brantford, CT, USA) instrument, capable of sequencing 25 million bases in a 4-h period, an advance 100-times more rapid than Sanger’s capillary-based electrophoresis method [16]. In this method, DNA is amplified using a ‘clonal’ approach and sequenced using a microfabricated, massively parallel platform. Adaptors are attached

to the sheared 300-bp genomic DNA fragments, permitting their capture on tiny beads (28 nm in diameter), and reaction conditions are optimized to favor the attachment of just one fragment per bead. Subsequently, oil droplets containing all the requisite reactants for DNA amplification encase the beads, forming an emulsion which maintains each bead distinct from its neighbor. This ensures uncontaminated amplification of approximately 10 million copies of the initial fragment. The beads are then dispensed into the open wells of a fiber optic slide and ‘pyrosequenced’ (which detects extension via luciferase-based, real-time monitoring of pyrophosphate release) [17,18]. This generates sequencing reads 100 base pairs in length. Shotgun sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome was carried out as an experimental case to highlight the throughput and accuracy of this approach. In a 4-h experiment using this system, 96% genome coverage was obtained with an accuracy of 99.96%.

Another approach for ultra-high-throughput DNA sequencing was reported by Shendure *et al.* [19]. This method of sequencing by synthesis on a solid support is similar in principle to that described by Margulies *et al.* [16]. This approach differed with regard to the method utilized for library construction, with respect to the sequencing chemistry, signal detection (which employs a modified epifluorescence microscope) and the array platform, but was particularly impressive in that the protocols were implemented with off-the-shelf instrumentation and reagents. The method relies on discrete clonal amplifications of a single DNA molecule that are grown on a solid-phase surface to give rise to polymerase colonies or polonies. A polony protocol is employed to generate a DNA library containing approximately 1.6 million fragments, each 135 bp in length with 100 bp in common, in addition to two ‘mate-pairs’ sequence tags 17 and 18 bp in length, respectively, which are derived from the genome being sequenced. The tags represent random sequences located approximately 1 kb apart on the genome. Each fragment is attached to a separate bead (1 µm in size), amplified using emulsion PCR, and immobilized in a polyacrylamide gel. Parallel sequencing is carried out using a four-dye ligation protocol to identify each base. For each fragment, a 26-bp sequence (13 bp from each tag) is determined. An *Escherichia coli* strain, MG1655, engineered for deficiencies in tryptophan biosynthesis was resequenced using this approach with an error rate estimated at one per million consensus bases.

Sequencing by synthesis as described above requires a nucleic acid amplification step, so that an adequate signal level can be achieved [20]. By contrast, single molecule approaches circumvent amplification and involve direct sequencing of single DNA molecules. One popular single molecule approach gaining in momentum is nanopore sequencing. As DNA passes through a 1.5-nm nanopore (a small pore in an electrically insulating membrane), different base pairs obstruct the pore to varying degrees, causing measurable variations in the electrical conductance of the pore, which can then be used to infer the DNA sequence [20,21].

Commercial high-throughput sequencing platforms

The first commercial next-generation sequencing platform was introduced in 2005 by 454 Life Sciences. The current Genome Sequencer FLX™ system from 454 Life Sciences, has reduced error rates and increased read lengths by 250 bp on average. Other commercial efforts include technologies from Illumina (San Diego, CA, USA), Applied Biosystems (Foster City, CA, USA), Helicos BioSciences (Cambridge, MA, USA) and Visigen Biotechnologies (Houston, TX, USA) amongst others.

The Illumina (Solexa) 1G Analyzer is based on the massively parallel sequencing of millions of fragments using a proprietary clonal single molecule array technology coupled to a novel reversible terminator-based sequencing chemistry. For short sequence reads, the approach has been determined to be highly robust and accurate. Applications in whole-genome association studies, expression analysis and sequencing, in addition to genome-wide location studies, have been described [5,22,23]. The short individual read lengths, comprising on average 25 bp in length, currently point to primary applications in re-sequencing where a known reference genome exists rather than sequencing *de novo*. Efforts are currently underway to develop algorithms and approaches for *de novo* assemblies using ‘paired end’ sequences.

The Applied Biosystems (Foster City, CA, USA) ‘supported oligo-ligation detection’ (SOLiD) DNA-sequencing system developed initially by Agencourt Personal Genomics, utilizes a related clonal amplification approach on beads, employing four fluorescent tags with a two-base readout system. Each ligation step interrogates a pair of adjacent nucleotides with each base interrogated twice for higher accuracy.

Applications include detection of sequence variation, such as SNPs, gene copy number variations, single base duplications, inversions, insertions and deletions.

Microarray-based genomic selection

The next generation of technologies is poised to reduce DNA sequencing costs by several orders of magnitude. This promises a more cost-effective and comprehensive determination of genetic variation and the healthcare potential for individualized patient genome sequencing. In order to fully leverage the power of this technology, obstacles in template preparation must be overcome. PCR, which has been the dominant enrichment technology for conventional sequencing, is rate limiting with newer technologies, as it requires the synthesis of large numbers of oligonucleotides and large numbers of individual reactions. Furthermore, PCR does not multiplex efficiently. Complex eukaryotic genomes are simply too large to explore without the use of complexity reduction methods. One such approach is microarray-based genomic selection (MGS), which permits enrichment of predefined sequences from complex eukaryotic genomes. Although several related techniques have been described, MGS essentially consists of shearing genomic DNA into smaller fragments that are ligated with unique adaptors, and subsequently hybridized to high-density oligonucleotide microarrays. The bound fragments are eluted and amplified by PCR using the adaptor primers and are subsequently sequenced [8–11].

‘The next generation of technologies is poised to reduce DNA sequencing costs by several orders of magnitude.’

Okou *et al.* reported MGS capable of enriching targeted sequences from complex eukaryotic genomes without the repeat blocking steps necessary for bacterial artificial chromosome-based genomic selection [8]. To test the approach, they captured and resequenced two X chromosome-linked genomic regions, initially a 50-kb region including coding and noncoding sequences surrounding the fragile X mental retardation 1 gene (*FMR1*), and later a larger-scale experiment that interrogated 304 kb of unique coding and noncoding sequences contained within a 1.7-Mb genomic region that includes *FMR1*, fragile X mental retardation 1 neighbor (*FMR1NB*) and AF4/FMR2 family, member 2 (*AFF2*). Custom oligonucleotide

microarrays from NimbleGen Systems, Inc. (Madison, WI, USA) containing 385,000 capture probes (50–93 bp in length) were designed to achieve optimal isothermal hybridization across the microarray. Resequencing was performed using a custom NimbleGen microarray.

‘Individual genome sequencing will find its niche in the area of personalized and preventative medicine.’

Hodges *et al.* focused on coding exons and their adjacent splice sites, a sequence range representing roughly 1% of the human genome [9]. Six custom NimbleGen arrays with overlapping 60–90-nt probes were designed, each containing 385,000 unique features that allow tiling of approximately 6 Mb of exonic sequence. An additional array designed to capture alternative transcripts facilitated the tiling of 37,000 exons. The enriched material was sequenced using an Illumina 1G Analyzer. Analysis of the captured fragments revealed that up to 85% derived from the targeted regions, and up to 98% of the expected exons were recovered. Albert *et al.* coupled high-density NimbleGen microarrays with 454 Life Sciences FLX sequencing to perform MGS [10]. A total of 6726 base exon segments approximately 500 bp in size and locus-specific regions 200 kb to 5 Mb in size were enriched and sequenced. The majority of the sequence reads represented selection targets.

Porreca *et al.* described an interesting variation of MGS that utilized a modification of the molecular inversion probe methodology to enrich sequences for Illumina 1G Analyzer sequencing. In this approach, 100-mer oligos are synthesized and released from a programmable microarray (Agilent Technologies, CA, USA), amplified using PCR and restriction digested to release a single-stranded 70-mer capture probe mixture. Each individual mature probe contains a universal 30-nt motif, flanked by unique targeting arms, each 20 nt in length. The arms hybridize immediately upstream and downstream of a specific genomic target, which is copied by polymerase-driven extension from the 3′ end of the capture probe. Ligation to the 5′ end completes a circle which is enriched and amplified. Advantages of this approach include compatibility with extensive multiplexing (facilitating capture of up to 10,000 targets in an individual reaction), high specificity with 98% of amplicons corresponding to targets, and the precise specification of target boundaries.

Implications for pharmacogenomics

The emergence and acceptance of high-throughput sequencing technologies will have consequences for the future use of microarrays. The full-genome arrays widely used today will eventually become defunct in favor of low-cost sequencing approaches that are not confounded by the innumerable problems that plague current microarray experiments, namely cross-hybridization of related species, poor hybridization kinetics, interference from DNA secondary structure, noise from repetitive elements, poor sensitivity in relation to low-abundance transcripts and the inability to distinguish between genes of interest and pseudogenes.

Many of the platforms in use today do not discriminate mRNA splice-variants, a limitation easily addressed by MGS coupled to high-throughput sequencing. Another major constraint with microarray analysis is that individual probes possess different temperatures for optimal binding to their complementary strands.

‘The full-genome arrays widely used today will eventually become defunct in favor of low-cost sequencing approaches that are not confounded by the innumerable problems that plague current microarray experiments.’

In recent years, a technique combining chromatin immunoprecipitation (ChIP) with microarray technology (chip) has been widely used to investigate *in vivo* interactions between proteins and DNA (ChIP–chip). A recent improvement to this technique (ChIPSeq) dispenses with microarrays and identifies protein-bound DNA fragments by direct DNA sequencing [22]. The main advantage of ChIPSeq over conventional ChIP–chip is that the entire genome can be assayed rather than the DNA regions captured by array probes [22–25].

Ultra-high-throughput technologies will find considerable use in resequencing genomes, genome wide epigenetic studies using sodium bisulfite sequencing, novel gene discovery efforts, such as uncovering small RNAs and microRNAs, sequencing of microbial genomes and pathogen identification. Expression profiling during the development of an organism from fertilized egg to maturity will become a routine exercise. The individual sequence information obtained using these technologies will serve as digital tags, providing a direct measurement of transcript copy number. This will be

particularly useful in the quantification of alternative splice variants in the transcriptomes of healthy and diseased cells.

Individual genome sequencing will find its niche in the area of personalized and preventative medicine. As costs currently remain too high to support routine resequencing of individual human genomes, personalized sequencing efforts will, in the short-term, focus on defined genomic regions that encompass disease-causing genes.

The era of 'retail genomics' is well underway with clinical diagnostic and prognostic testing routinely available for common and rare inherited disease conditions, risk assessment and prevention, and *a priori* stratification for pharmacogenetic contraindications. As this evolves from genotyping to complete genome sequencing, ethical issues currently under consideration will become even more pertinent. A scenario where a patient opts to remain ignorant of a predisposition to a late-onset disease, particularly one that cannot be treated, prevented or ameliorated is

understandable. That public disclosure of such information could influence health insurability and employment prospects is a frightening possibility. Greater pressure to live healthy lives, predictive diagnoses and treatment long before any physical evidence of disease manifests will exploit the use of personalized genome-sequencing methods. Appropriate retraining of medical personnel in genomic medicine will be an *a priori* requirement so that they are better equipped to counsel and treat patients presenting with the awareness of possessing genetic aberrations.

Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending or royalties. No writing assistance was utilized in the production of this manuscript.

Bibliography

- Marton MJ, DeRisi JL, Bennett HA *et al.*: Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* 4, 1293–1301 (1998).
- Brown PO, Botstein D: Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21, 33–37 (1999).
- Hardiman G: Microarray platforms – comparisons and contrasts. *Pharmacogenomics* 5, 487–502 (2004).
- Hardiman G, Carmen A: DNA biochips – past, present and future; an overview. In: *Biochips as Pathways to Discovery*; Carmen A and Hardiman G (Eds). Taylor & Francis, NY, USA 1–13 (2006).
- Fields S: Molecular biology. Site-seeing by sequencing. *Science* 316, 1441–1442 (2007).
- Zwick ME: A genome sequencing center in every lab. *Eur. J. Hum. Genet.* 11, 1167–1168 (2005).
- Rogers YH, Venter JC: Genomics: massively parallel sequencing. *Nature* 437(7057), 326–327 (2005).
- Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* 11, 907–909 (2007).
- Hodges E, Xuan Z, Baliya V *et al.*: Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527 (2007).
- Albert TJ, Molla MN, Muzny DM *et al.*: Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 11, 903–905 (2007).
- Porreca GJ, Zhang K, Li JB *et al.*: Multiplex amplification of large sets of human exons. *Nat. Methods* 11, 931–936 (2007).
- Sanger F, Coulson AR: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94, 441–448 (1975).
- Sanger F, Air GM, Barrell BG *et al.*: Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687–695 (1977).
- Maxam AM, Gilbert W: A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* 74, 560–564 (1977).
- Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463–5467 (1977).
- Margulies M, Egholm M, Altman WE *et al.*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380 (2005).
- Ronaghi M: Pyrosequencing sheds light on DNA sequencing. *Genome Res.* 11, 3–11 (2001).
- Gharizadeh B, Nordström T, Ahmadian A, Ronaghi M, Nyrén P: Long-read pyrosequencing using pure 2'-deoxyadenosine-5'-O⁻-(1-thiotriphosphate) Sp-isomer. *Anal. Biochem.* 301, 82–90 (2002).
- Shendure J, Mitra RD, Varma C, Church GM: Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 5, 335–344 (2004).
- Shendure J, Porreca GJ, Reppas NB *et al.*: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732 (2005).
- Winters-Hilt S, Vercoutere W, DeGuzman VS, Deamer D, Akeson M, Haussler D: Highly accurate classification of Watson–Crick basepairs on termini of single DNA molecules. *Biophys. J.* 84, 967–976 (2003).
- Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* 316, 1497–1502 (2007).
- Barski A, Cuddapah S, Cui K *et al.*: High-resolution profiling of histone methylations in the human genome. *Cell* 129(4), 823–837 (2007).
- Euskirchen GM, Rozowsky JS, Wei CL *et al.*: Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res.* 17, 898–909 (2007).
- Robertson G, Hirst M, Bainbridge M *et al.*: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657 (2007).